III Conferecia Bibliotecas y Repositorios Digitales de América Latina (BIREDIAL '13)
VIII Simposio Internacional de Bibliotecas Digitales (SIBD '13)
"ACCESO ABIERTO, PRESERVACIÓN DIGITAL Y DATOS CIENTÍFICOS"
Ciudad de la Investigación, Universidad de Costa Rica, del 15 al 17 de octubre de 2013.
Ponencia

Control de Autoridades en DSpace

Sergio NIETO, Emilio LORENZO

Arvo Consultores

Resumen: El control de autoridades es una de las piezas clave puestas que tienen los responsables de Repositorios Digitales para mejorar la calidad de contenidos y posibilitar la interoperabilidad entre repositorios. En este texto revisamos los mecanismos de tipo esquema semántico, entre los que se encuentra el control de autoridades, evaluaremos cómo diversos sistemas federados de contenidos incorporan estos mecanismos para mejorar la interoperabilidad y finalmente revisaremos alguna de las opciones disponibles para el uso de control de autoridades de nombre de autor. En el taller adjunto examinaremos algunas de las herramientas y modos de uso del módulo de control de autoridades que incorpora DSpace.

Palabras Clave: Repositorios Digitales, DSpace, authority control, control de autoridades

Introducción

Los Repositorios Institucionales se diseñaron con la idea de convertir el auto-archivo en la estrategia principal de depósito de objetos digitales, pero desafortunadamente el auto-archivo lleva parejo el uso de relativamente pocos elementos para controlar la calidad de los metadatos. El difícil equilibrio entre efectuar un descripción precisa de los objetos digitales o simplificar el proceso de depósito a costa de relajar controles y simplificar pantallas, normalmente se decanta por esta segunda opción. Por ello, la calidad de los metadatos de los repositorios es manifiestamente mejorable, tal y como nos recuerdan a menudo algunos estudios[i].

Entre los elementos adicionales que los Repositorios Institucionales (junto con otros sistemas de información relacionados) ofrecen a los bibliotecarios y técnicos responsables del repositorio para mejorar la calidad de los metadatos figuran los vocabularios controlados, los tesauros, las ontologías y el control de autoridades.

El control de autoridades para nombre de autor es una de las piezas clave que disponemos para la mejorar la calidad de contenidos y para mejorar la interoperabilidad entre repositorios. En este taller examinaremos algunas de las herramientas y técnicas implícitas en el framework de **authority control** de DSpace para lograr esta mejora.

Interoperabilidad y control de autoridades

La racionalidad del control de autoridades deriva principalmente de tres consideraciones: 1) la necesidades de mejorar la interoperabilidad entre sistemas, 2) proporcionar medios por los que las personas y los sistemas de información pueden comunicarse con la menor ambigüedad posible 3) mejorar las posibilidades en que diferentes agentes puedan llegar a la misma representación cognitiva.

Se entiende por interoperabilidad *la capacidad de los sistemas o componentes de intercambiar*

información y de poder controlar el procesamiento cooperativo entre aplicaciones[ii]. La interoperabilidad puede ser analizada desde distintos niveles, tal y como lo especifica la Dublin Core Metadata Initiative, DCMI[iii], Los niveles semánticos 1 y 2, que en terminología DCMI, incluirían a Shared Term Definitions y Formal Semantic Interoperability, son de nuestro mayor interés pues se definen como la capacidad de los sistemas de información para intercambiar información basándose en un significado común de los términos contenidos en los metadatos y documentos. El objetivo perseguido es asegurar la consistencia, representación y recuperación de los contenidos de nuestro repositorio.

Entre los elementos tecnológicos que nos ayudan a alcanzar estos objetivos, aparecen de forma principal los esquemas semánticos, y bajo este término se incluyen vocabularios controlados, tesauros, listados de encabezamiento de materias, anillos de sinónimos, taxonomías, ontologías, etc. Los esquemas semánticos proporcionan elementos efectivos (aunque no siempre de fácil uso) por los cuáles las personas interrelacionándose e interactuando con/a través de sistemas de información pueden reducir la ambigüedad de la comunicación de conceptos.

Vocabularios controlados

Un vocabulario controlado es un conjunto de unidades léxicas seleccionadas con fines de efectuar una normalización terminológica.

Tesauros

"Se trata de un vocabulario controlado y *estructurado* al que se llega mediante la selección de términos del lenguaje natural. Por lo tanto , está constituido por una lista de palabras, llamadas descriptores, estructuradas de forma que unas se relacionan con otras". [iv]

Un Tesauro es una lista que contiene los «términos» empleados para representar los conceptos, temas o contenidos de los documentos, interrelacionados entre ellos bajo un conjunto de modalidades de relación, entre las que destacan:

1. *Relaciones jerárquicas*: Establecen subdivisiones que generalmente reflejan estructuras de un todo y sus partes...,
2. *Relaciones de equivalencia*: Controlan la sinonimia, homonimia, antonimia y polisemia entre términos.
3. *Relaciones asociativas*: Mejoran las estrategias de recuperación y ayudan a reducir la polijerarquía entre los términos.

Ontologías

Son modelos formales describiendo cómo percibimos los estados posibles de un dominio de conocimiento. Se pueden considerar los sucesores tecnológicos de los tesauros, y a menudo su distinción es dificultosa. Mientras los tesauros identifican el significado y relaciones de un conjunto de conceptos, las ontologías se enfocan en analizar los marcos de pensamiento sobre esos conceptos y tratar de codificarlos en un lenguaje formal[v].

Control de Autoridades

Una **autoridad** es un conjunto de valores controlados para un dominio determinado, estando cada valor único identificado por una clave (clave de autoridad). Un **registro de autoridad** es la información asociada con cada uno de los valores en una autoridad (incluyendo variaciones de deletreo, valores equivalentes y/o alternativos, etc). Una **clave de autoridad** es un identificador opaco y persistente correspondiente a un registro de autoridad.

En la práctica habitual, un registro de autoridad (de nombres de autor, por ejemplo), contiene la forma autorizada del nombre del autor, establecida por la institución normalizadora como forma preferida para visualizar en sus sistemas, así como las formas variantes del nombre y nombres relacionados. Además, el registro de autoridad puede contener información relativa a la persona, representada por el punto de acceso), así como a las relaciones entre esa persona y otras entidades relacionadas, información para identificar las reglas de acuerdo con que se establecieron valores controlados, las fuentes consultadas, la agencia de catalogación encargada de establecer la normalización y la agencia responsable de establecer las formas preferidas del nombre.

Entre los **objetivos** que un control de autoridades busca, podemos relacionar:

- Dar consistencia e integridad a los metadatos, ayudando en la corrección de los correspondientes valores . En un repositorio digital esto se consigue conectando las interfaces de archivo y edición con el sistema de autoridades con el fin de chequear los valores introducidos contra los registros de autoridad.
- Conseguir mejorar la precisión en la recuperación de la información, puesto que el mejor método, simple y positivo, de determinar si dos valores son idénticos, es comparando las claves de autoridad, ya que comparar valores textuales proporciona falsos positivos (demasiados García, M.) o falsos negativos (¿García, M. vs. García, Manuel?)
- Facilitar el intercambio de información bibliográfica, ya que al compartir autoridades el repositorio puede interoperar de manera más efectiva. Por ejemplo el uso de identificadores normalizado de autor, facilita enormemente la incorporación de publicaciones al repositorio

Interoperabilidad semántica en sistemas federados de repositorios

Con la finalidad de integrar y consolidar las producciones científicas y académicas existentes en múltiples repositorios dispersos, aparecen sistemas federados de repositorios, impulsados a nivel nacional, internacional o incluso disciplinar. El mecanismo de interoperabilidad técnica-sintáctica que se usa como estándar de facto entre repositorios es el intercambio de metadatos OAI-PMH, que conecta a los **proveedores de datos** (repositorios digitales) con los **proveedores de servicios** (agregadores, cosechadores o recolectores.)

Lamentablemente OAI-PMH no incorpora mecanismos para lograr los objetivos buscados por la interoperabilidad semántica. Aunque el protocolo permite el intercambio usando esquemas de metadatos distintos del dublin-core, pocos repositorios y pocos agregadores - recolectadores van más alla de este esquema. En concreto, el espacio de nombre dublin-core no ofrece posibilidad prácticas de distinguir un nombre autorizado o controlado, ni , por ejemplo, de reflejar la afiliación de un autor a una determinada institución. Hay más problemas que inciden en la calidad de los metadatos en sistemas federados, pues el sistema recolector no recibe ninguna indicación del grado de precisión de los mismos, sólo puede asumir que todos son igualmente válidos. Igualmente no hay una manera de comunicación de correcciones de los repositorios a los agregadores, dando lugar en ocasiones a registros duplicados con pequeñas variaciones entre los metadatos, sin ninguna indicación de la preferencia de un registro sobre otro....

Una revisión inevitablemente resumida de cómo los proveedores de servicios abordan, es decir qué especificaciones imponen a los proveedores de datos, nos conduce a la siguiente tabla resumen:

	Interoperabilidad Técnica/sintáctica	Interoperabilidad Semántica

REMERI, Red mexicana repositorios institucionales (MEX)	OAI-PMH con dublin-core Se recomienda el cumplimiento con directrices Driver 2 (Digital Repository Infrastructure Vision for European Research)[vi]	Se recomienda el cumplimiento con directrices Driver 2. Vocabularios controlados en una variedad de campos
BDCOL, BIBLIOTECA DIGITAL COLOMBIANA (COL)	OAI-PMH con dublin-core, aunque se recomienda el uso del esquema específico BDCOL (namespace http://www.bdcol.org/documents/metadata/)	Vocabularios controlados en una variedad de campos, entre ellos los correspondientes al esquema ETD-MS
RECOLECTA , recolector de ciencia abierta, ESP	OAI-PMH con dublin-core. planificada la compatibilidad sintáctica mediante directrices Driver2 a finales de 2013	Planificada la compatibilidad semántica (vocabularios controlados) mediante directrices Driver2 a finales de 2013
EUROPEANA, biblioteca digital europea de acceso libre, EU	OAI-PMH con dublin-core y europeana semantic elements (15 metadatos Dublin Core, un subgrupo de DC terms y 13 elementos específicos europeana)	Europeana semantic elements, prevista su evolución a Europeana Data Model
OPENAIRE, Open Access Infrastructure for research in Europe, EU	OAI-PMH con dublin-core Se requiere el cumplimiento con directrices Driver 2 (Digital Repository Infrastructure Vision for European Research).	Vocabularios Driver 2 y subconjunto de vocabularios OpenAire Driver2: Vocabularios controlados y sintaxis recomendada de campos no controlados (como autor) posibilidad de incorporar valores de autoridad mediante el uso de namespaces adicionales (responsabilidad del proveedor de datos

Resumiendo, los proveedores de servicio intentan abordar el problema de la interoperabilidad, principalmente mediante un doble enfoque:

- Especificación de los esquemas de metadatos admisibles y de los sub-perfiles (campos obligatorios y opcionales) de los esquemas empleados (es decir, interoperabilidad sintáctica), por ejemplo haciendo obligatorios el uso de determinados campos de un esquema (dc.contributor, dc.format, dc.language...)

- Especificación de vocabularios controlados para determinados elementos del esquema de metadatos (interoperabilidad semántica), forzando la normalización de los datos contenidos en determinados campos, en general, de aquéllos de uso obligatorio (dc.type, dc.language)

y marginalmente, como planteamiento adicional, la especificación Driver 2 abre la posibilidad de incluir valores de autoridad, siempre y cuando la parte autoritativa que actúa como Agencia de registro de autoridades sea incluida y pueda ser reconocida en el esquema. En este sentido, la práctica recomendada consiste en que los Proveedores de Datos codifiquen la clasificación de autoridad de forma “URI-ficada” utilizando el espacio de nombres de autoridad para respaldar el reconocimiento del esquema y por ende de los valores de autoridad empleados.

Sistemas de autoridades para nombres de autor

Una vez revisada la situación real del uso de autoridades en los repositorios, y centrándonos en

las autoridades para nombre de autor ¿cuáles son las opciones más viables y accesibles para empezar a caminar en la implantación de un sistema de autoridades? Podemos considerar dos posibilidades:

- **Uso de Proveedores externos de información**
- **Uso de información institucional como fuente de nombres de autoridad**

Uso de Proveedores externos

Diremos en primer lugar que los repositorios están aún técnicamente poco preparados para la interconexión con los diversos servicios de consulta disponibles, ya sea porque no son tantos los servicios que ofrecen formas convenientes, técnicamente hablando, de integración, (OCLC, la Biblioteca del Congreso Estadounidense, etc,) como por el hecho de que normalmente se requieren desarrollos y programaciones específicos en el software del repositorio para acceder a las pocas interfaces de trabajo (consulta, recuperación, validación...) ofrecidas por los proveedores.

En segundo lugar, algunos de estos servicios son declaradamente sectoriales, cubriendo autores en determinada disciplina o de una determinada geografía... El resto de autores fuera de ese ámbito declarado, no tendrán registro de autoridad. La amplia cobertura disciplinar o universalidad de la mayoría de los repositorios digitales colisiona con esta característica de la mayoría de proveedores de registros de autoridad. La aparición de proveedores con vocación multidisciplinar y multinacional previsiblemente cambiará el panorama a medio plazo.

Nuestra recomendación es que el uso de proveedores externos como *única* fuente de registros de autoridad no es aún una opción viable en el corto plazo. No obstante proveemos una relación, de nuevo limitada, de proveedores de sistemas de autoridad de nombres, servicios de desambigüación de nombres o registros de identificadores únicos de autor:

	Ámbito	
International Standard Name Identifier (ISNI) [1]	Identificación Única de Identidades Públicas de múltiples campos de la actividad creativa (sector de medios, cadenas de creación, producción, gestión y distribución de contenidos)	Desarrollado por la International Standards Organization, pretende convertirse en un estándar, basado en la Norma ISO 27729
VIAF, Virtual International Authority File, [2]	Asigna identificadores en el ámbito de Bibliotecas Nacionales principalmente. Library of Congress, BNF, BN República Checa, BN Alemania, BN Israel, BN Suecia, BN Australia, BN España, BN Portugal, ICCU (Italia)..	Crea una red virtual de registros de autoridad de nombres de persona
IraLIS, International Registry of Authors-Links to Identify Scientists[3]	Autores principalmente de habla hispana	Es la autoridad de nombres del repositorio E-LIS
ORCID, Open Researcher and Contributor ID [4]	Registro abierto de identificadores de investigadores y autores, con enlace a sus publicaciones.	Pretende convertirse en un estándar para la identificación única y persistente de autores y posiblemente su histórico de afiliación institucional

[1] http://www.isni.org/
[2] http://www.oclc.org/research/projects/viaf
[3] http://www.iralis.org/
[4] http://orcid.org/

RePEc [5]	Investigación en el área de Economía, Historia Económica y Empresa y Estadística	El servicio RePec Authors Service permite el enlace de registros de autor con la producción científica que haya sido depositada en la base de datos RePeC.
ResearcherID (Thomson Reuters) [6]	Asignar identificadores únicos a autores referenciados en WoK	Dispone de integraciones para que los repositorios puedan descargar las publicaciones de un Identificador, así como la integración con ORCID
Pilot National Name y Factual Authority Service, JISC	Investigadores en instituciones del Reino Unido	Continuación del proyecto Names [7]. Actualmente el servicio de autoridades de nombres se encuentra en fase de prototipo
SURFfoundation[8], Digital Author Identifier	Investigadores en instituciones holandesas	Conectado con el National Thesaurus of Authors' Names (NTA) provee a los autores de un identificador compatible con el ISNI
FRIDA (Noruega); Researcher Name Resolver (Japón); DISSOnline (Alemania)	Autores o partícipes en sus respectivos sistemas nacionales de investigación	

Como ya apuntamos, los repositorios que evalúen fuentes externas para el control de autoridades, se encontrarán con una panorama confuso, en el que una diversidad de proveedores compiten intentando alcanzar la masa crítica que hará realmente valiosos sus servicios. Los servicios nacionales de autoridad tendrán que competir (en caso de existir) con servicios comerciales con un indudable atractivo para los investigadores y en esta proliferación, los responsables de repositorios normalmente acaban implantado soluciones ad-hoc, recurriendo a fuentes internas funcionando como nombres de autoridad.

Uso de información institucional como fuente de nombres de autoridad

Las ventajas son evidentes, pero también debieran ser los inconvenientes. La mayor ventaja es que la disponibilidad de los nombres e identificadores únicos de los autores pertenecientes a la institución simplifica el uso de esos datos para asignar valores únicos a los autores "propios" sin necesidad de recurrir a servicios externos. Por contra, como principal desventaja, los procesos de desambigüación de nombres y asignación de un valor de autoridad suele ser un proceso intensivo en recursos expertos, que escasean siempre en una mayoría de repositorios. Además el modelo "básico" de autoridades que proporciona DSpace y que describiremos a continuación tiene limitaciones en lo referente a contemplar variaciones de nombres, homogeneización de firmas, etc

[5] https://authors.repec.org/about
[6] http://www.researcherid.com/
[7] http://names.mimas.ac.uk/
[8] http://www.surf.nl/en/themas/openonderzoek/infrastructuur/Pages/digitalauthoridentifierdai.aspx

El modelo de authority control de DSPACE

El modelo de autoridades de Dspace aparece en la versión 1.6 del aplicativo de forma estándar, pues previamente era una pieza de código separada como add-on.

Se define como un framework que mediante configuración permite conectar (plug-in) clases programáticas para controlar dos aspectos básicos: Cómo se realiza la selección de valores en un metadato **(choice management)** y la inclusión de valores de autoridad asociados a los valores de metadatos **(Authority Control).**

Por tanto, **no ofrece** funcionalidad alguna para la gestión de las autoridades, de los registros de autoridad o de las claves de autoridad, que se consideran fuera del ámbito de DSpace.

Junto con el código que ofrece la funcionalidad de control de autoridades, se distribuyen con DSpace una serie de conectores con servicios de autoridades ya existentes, a modo demostrativo, como el Servicio de Nombres de la Biblioteca del Congreso (Library of Congress Names service[9]), y el servicio de autoridades de nombres de revistas y editores Sherpa-Romeo [10]

Funcionalidades básicas del framework

Choice Management

En los elementos del Interfaz de usuario que se ocupan de la edición de metadatos (principalmente módulo de envíos para usuarios de autoarchivo u módulo de edición de metadatos para administradores) se pueden incluir funcionalidades que asisten en la selección de valores de los metadatos que se hayan configurado. Para dichos campos de metadatos se pueden generar listas de valores a partir de vocabularios extensos, navegación por tesauros jerárquicos, selección cerrada a los valores de una lista, lista abierta, …

Authority Control

El control de autoridades proporciona incluye la clave de autoridad junto con el valor del metadato seleccionado. Señalar que los metadatos controlados por autoridad deben llevar asociado el plugin Choice management.

La información de autoridad consiste del valor del metadato, el valor de la clave de autoridad (authority key) y el denominado valor de confianza (confidence value), cuya utilidad explicaremos más adelante.

Mientras que el uso de vocabularios controlados sólo es relevante en la interfaz de archivo, el control de autoridades se usa siempre que un metadato se cambia, y esto incluye el archivo de ítems vía batch, los envíos sword, las interfaces administrativas, etc..

Visibilidad de las claves de autoridad y de los valores de confianza

En la interfaz OAI_PMH se expone únicamente el valor del metadato, (ya mencionamos antes la limitación del protocolo para exponer valores de autoridad) estando ocultos los valores de autoridad y de confianza, mientras que en las interfaces de usuario, se exponen, y por tanto pueden usarse para ofrecer información adicional al usuario. Por esta misma razón se recomienda el uso de claves de autoridad que no ofrezcan información personal de autores, como pudiera ser cuentas de correo, u otros códigos de identificación.

[9] http://id.loc.gov/authorities/names.html
[10] http://www.sherpa.ac.uk/romeo/

Indices de autoridad

Una característica normalmente poco conocida es la posibilidad de construir índices (browse o search indexes) que contengan sólo valores con clave de autoridad asociada. Así podemos tener un índice de autores y otro índice que incluya sólo autores validados. Además, la inclusión de un valor validado en este índice puede ser controlada mediante el valor de confianza.

El valor de confianza (confidence value) se expresa como un valor simbólico dentro del rango siguiente: (aceptado, incierto, ambiguo, no encontrado, fallido) y puede asignarse adicionalmente al valor de clave de autoridad. A continuación, podemos especificar el nivel inferior de confianza que es necesario para incluir un valor de metadato en el índice construido, con lo que el índice así construido, incluirá los valores validados con ciertas condiciones, que sobrepasen ese nivel inferior (minimun confidence value).

Los registros de autoridad

Los registros de autoridad son externos a DSpace, es decir, DSpace no incluye ninguna funcionalidad para gestionarlos, depurarlos o ampliarlos, es decir, no incluye la posibilidad de añadir o asociar un valor adicional a un registro de autoridad ya existente.

Típicamente es una base de datos de la institución, un proveedor externo (como los relacionados en apartado anterior), un servidor de vocabularios, etc... La arquitectura de plugins de DSpace permite integrar conectores a estos servicios de forma simple, sin tocar el código original de DSpace.

El framework de DSpace incluye los conectores para tres servicios-proveedores ya mencionados, , el servicio de Nombres de la Biblioteca del Congreso y los servicios de autoridades Sherpa-Romeo de nombres de revistas y editores

org.dspace.content.authority.LCNameAuthority

org.dspace.content.authority.SHERPARoMEOPublisher

org.dspace.content.authority.SHERPARoMEOJournalTitle

y el conector que deberemos modificar y extender para conectarlo a servicios que desarrollemos:

org.dspace.content.authority.SampleAuthority

Configuración del Authority Control

Los subsistemas de Choice Management y Authority Control son un framework que debe ser configurado para que funcione. El comportamiento de cada campo de metadatos se configura exclusivamente con dos controles: Las propiedades en la sección correspondiente de dspace.cfg y el tipo de plugin ChoiceAuthority que hayamos seleccionado.

Para cada campo de metadatos se configuran los siguientes aspectos:

1. Activación para ese campo del plugin ChoiceAuthority, que pasa a controlar a partir de es momento el comportamiento del campo.
2. Estilo de presentación.
3. Selección de valores abierta o cerrada.
4. Control de autoridades

5. Valor de autoridad requerido.

La segunda parte de la intervención, en formato taller de trabajo, trata de la configuración del modelo de autoridades (Choice management y Authority Control), revisando los ficheros de configuración pertinente, así como los principales parámetros de configuración requeridos para incorporar esta funcionalidad a las pantallas de envío y edición de ítems.

Creación de authority-control de nombre de autores.

En la tercera parte del este taller, es decir con una vocación eminentemente práctica, se revisarán los pasos y técnicas necesarias para crear nuestro propio authority-control de nombre de autores. Para ello, se trabajará en el código y configuraciones requeridas para que los formularios de envío de ítems y los formularios de envío de edición de ítems funcionen sobre una base de datos de autores, ejemplificando alguna de las existentes en nuestras Instituciones, y permitan la selección y validación de los autores.

---

[i] Medrano, José Federico; Figuerola, Carlos G.; Alonso Berrocal, José Luis. Repositorios digitales en España y calidad de metadatos. // Scire. 18:2 (jul.-dic. 2012) 109-121. ISSN 1135-3716.

[ii] ISO19101:2002 International Standard Organization

[iii] Interoperability Levels for Dublin Core Metadata recuperado el 05/06/2013, de http://dublincore.org/documents/interoperability-levels/

[iv] Blanca Gil Urdiciain; Manual de Ciencias de la Documentación; Piramide, 2002

[v] Martin Doerr, (January 2008), "Ontologies", DCC Digital Curation Manual, S.Ross, M.Day (eds), recuperado 20/04/2013, de http://www.dcc.ac.uk/resource/curation-manual/chapters/ontologies

[vi] DRIVER 2.0. (2008). "Directrices para proveedores de contenido Exposición de recursos textuales con el protocolo OAI - PMH", DRIVER guidelines for Repository Managers translated into Spanish by RECOLECTA (2008). http://www.driversupport.eu/managers.htm